Association for Information Systems
AIS Electronic Library (AISeL)

SAIS 2015 Proceedings | Southern (SAIS)

2015

User Generated Content In Social Media As A Source For Assessing Cultural Dimensions

Geoffrey Hill
Kent State University, ghill11@kent.edu

USER GENERATED CONTENT IN SOCIAL MEDIA AS A SOURCE FOR ASSESSING CULTURAL DIMENSIONS

Geoffrey Hill
Kent State University
ghill11@kent.edu

ABSTRACT

This research in progress intends to document the process of mining social media text-based content in order to acquire measures of cultural dimensions. This process can potentially be replicated and extended to other topics of inquiry so as to provide researchers with an alternative source for data acquisition related to measuring theoretical constructs. Additionally, we provide an argument supporting the credibility of user generated content (UGC) as a data source for rigorous inquiry. This article describes the mechanism for acquiring pertinent UGC as well as describing the method of assessing and quantifying the valence and magnitude of the various construct values. These measures are intended to be suitable for utilization across a variety of methods for statistical analysis requiring continuous or discrete factor measures. We conclude with a discussion of limitations and benefits and intend to present preliminary results during the conference.

Keywords

Social media, unstructured text mining, automated classifiers, sentiment analysis, cultural dimensions

INTRODUCTION

Web 2.0 technologies such as social media have experienced tremendous amounts of popularity and growth not only in the amount of generated content but in the number and demographic variety of its users. YouTube claims in excess of one billion unique monthly users viewing over six billion hours of video and creating over 100 hours of video every minute (YouTube, 2014). Twitter claims 284 million active monthly users writing over 500 million messages every day (Twitter, 2014) with a record of 143,199 user generated messages in a single second (Krikorian, 2013). This combination of active users and spontaneously generated content could potentially provide a tremendously valuable source of data for academic inquiry.

The problem relates to harnessing these technologies for academic inquiry and simultaneously trusting the validity of the results. This research intends to address both portions of this question by detailing a mechanism that researchers can use to acquire data for assessing Hofstede's cultural dimensions (Hofstede, Hofstede and Minkov, 2010) and secondarily, we present an argument intended to bolster the validity ascribed to any derived results. These processes can then be modified to other focal area(s) of interest. The collection process will be described for the Twitter micro-blogging web service that provides an easy to use interface based upon industry standard web service communication mechanisms such as Representational State Transfer (REST) (Fielding, 2000) and JavaScript Object Notation (JSON)(Crockford, 2006).

Uni-grams	Bi-grams	Tri-grams
Shall	United States	President pro tempore
States	Vice President	Government United States
President	Shall power	Speaker house representatives
United	Which shall	Shall power enforce
State	Senate shall	United States shall
Such	Congress may	Temporary appointments until
Congress	President shall	Such number majority

Table 1. Keyword Extraction of the U.S. Constitution

The extraction of actionable knowledge from streams of unstructured text messages is a difficult and complex problem that has enjoyed limited success. Current mechanisms for deriving insights from such sources include topical trend analysis via keyword extraction, for example Paul and Dredze's (2011) inquiry of healthcare related topics in social media. Unfortunately, this mechanism provides only a limited perspective of the vast amount of content contained in social media streams. Keyword extraction is only able to provide a representation of the most used words within a text corpus (see Table 1). As such, its outputs are little more than a frequency representation of individual words (uni-grams) or word groups (n-grams) that appear within the corpus under investigation.

Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan, 2003), another popular method of representing information in text streams, provides a visualization mechanism of topic trends usually expressed as word clouds (see Figure 1). LDA extends upon the simplistic keyword extraction method by aggregating words together into common factors based upon "a probabilistic model for collections of discrete data. LDA is based on a simple exchangeability assumption for the words and topics in a document…" (Blei et al., 2003). However, it provides no mechanism by which the individual author's magnitude and valence of sentiment expression can be assessed regarding any given factor. Thus, these current methods serve only to provide aggregate representations of the themes associated with a given corpus as a clustering technique of data compression.

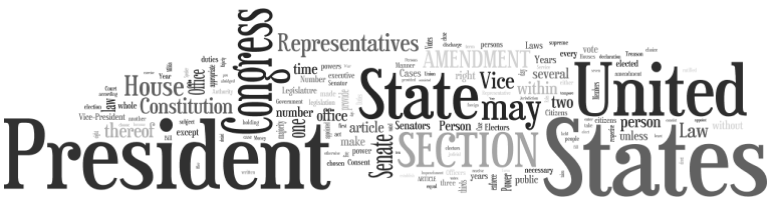

Figure 1. Word-Cloud of the U.S. Constitution

Contrasting these methods of analyzing text streams, we intend to provide a method by which each individual message can be assessed against a set of relevant factors as an individualized response. This expands upon the capability to extract knowledge from social media streams by treating authors and messages individually instead of in aggregate. Our method assesses each message as a singleton which can then be fused together at higher levels of aggregation (e.g. author, topic, etc.) based upon the needs of the research instead of being dictated by limitations inherent to the analytical method. This will allow for the individualized measurement of each participant within social media streams across a variety of factors as determined by a hypothesized measurement model. These individualized measures can then be subjected to any of a variety of statistical methods for example regression, clustering, and structural equation modeling.

THEORETICAL BACKGROUND

Before any data can be subjected to analysis, we must be able to trust that the data is valid, correct and appropriate. Information adoption theory (Sussman and Siegal, 2003) indicates that information will be adopted, or accepted as valid, based upon the quality of an underlying argument as well as the source credibility of the originator. For the purposes of this research, the quality of the underlying argument is achieved in the originating theoretical work describing the focal constructs. At this time, we are not intending to propose any new models or rehash theoretical models in use across our discipline. We intend simply to provide an alternative mechanism for measuring their constructs. Therefore, this research will stipulate as to the quality of the originating theories describing the constructs we use here. This allows us to then focus upon establishing the second portion specified by information adoption theory; credibility.

Credibility is a well explored phenomenon spanning multiple disciplines. The credibility of threats are often described in intelligence and law-enforcement literature as a multi-dimensional construct comprised of opportunity, capability and intent (Rynes and Bjornard, 2011). Although these dimensions are of tremendous importance, the ability to accurately assess each one in isolation is limited in many web 2.0 environments such as social media. Social media messages are often characterized by their brevity and concision, particularly because some services place character limits upon message length. Thus, the volume of data contained in each message may not be comprised of enough information to accurately assess each and every requisite dimension. Therefore, the ability to accurately assess credibility as described by these works is limited at best.

Credibility theory (Longley-Cook, 1962) describes a process by which the credibility of a sufficiently large dataset may be measured for comparative purposes between multiple data sets and is largely impacted by the underlying size of the dataset. As the size of a dataset increases, so too does its credibility relative to other datasets. Thus, credibility is expressed as a characteristic of the dataset as a whole, and not as a characteristic of the individual data points. In order to fulfill our goal of being able to assess individual social media messages, this measure of credibility falls at an incorrect unit of analysis. We must identify a mechanism by which credibility can be established not at the dataset, but at the individual message.

Establishing credibility at the level of an individual social media message is not a novel concept and has seen some degree of academic interest as of late. Ha and Ahn (2011) described an investigation into how the presence of external links – uniform resource locators (URL) – impact upon users' credibility assessments of social media messages and found that the presence of URLs in a Twitter message (aka "tweet") positively influenced users credibility assessments of the tweets. Additionally, credibility of social media site contents has been found to be positively associated with the Unified Theory of Acceptance and Use of Technology (UTAUT) (Venkatesh, Morris, Davis and Davis, 2003) constructs in an investigation of gender

differences in using social media to acquire information relating to nonprofit organizations (Curtis, Edwards, Fraser, Gudelsky, Holmquist, Thornton and Sweetser, 2010). Thus, assessing message credibility is a topic of contemporary interest.

However, contemporary investigations into credibility in social media contexts are centered upon quantitatively assessing credibility. We contend that a viewpoint exists that may allow an assumption of credibility relating to individual messages. This viewpoint is in keeping with the traditions of jurisprudence regarding hearsay and its exceptions. The U.S. Federal Rules of Evidence (Fed. R. Evid. 803) allow for an exception to be granted and statements to be assumed as factual and entered into evidence where the statements can be characterized as one of the following: present sense impression, excited utterance, and then-existing mental condition. Each of these three characteristics can apply to social media messages.

Present sense impression describes or explains an event or condition, made while or immediately after the declarant perceived it (Fed. R. Evid. 803). An excited utterance relates to a startling event or condition, made while the declarant was under the stress of excitement that it caused (Fed. R. Evid. 803). A then-existing mental condition is a statement of the declarant's then-existing state of mind or emotional condition (Fed. R. Evid. 803). In today's commercialized world consumers are continually enticed to engage across social media and provide reactions to companies, organizations, special interest groups, etc. These prompts are often presented as links to social media sites or hashtag (#) keywords intending to evoke responses precisely characterized by these three descriptions associated with exceptions to hearsay. Some social media services, such as Twitter, may amplify these effects due to the inherent nature of micro-blogging and the exchange of concise messages often limited to a single sentence, thought, or fragment. Therefore, we contend that social media messages associated with micro-blogging services may be granted an expectation of credibility due to the mechanisms similar to those recognized by jurisprudence and the likelihood for such utterances to be free from guile and dissimulation.

We do not intend that this contention be interpreted such that each and every message is guaranteed to be credible. Instead, we are intending that this be interpreted as a likely probability of credibility. This likely probability expectation is derived from the tremendous volume, velocity and variety of messages associated with social media which has led to its classification as big data as described by Laney (2001). These characteristics allow us to extend our expectation of credibility to individual messages and contend that although individual messages may in fact be incredible; the previously described mechanisms applied to a big data environment indicate that any given message is likely to be credible due to the tremendous amounts of messages involved. In short, the expectation is that the incredible messages will be the outliers.

Apart from seeking mechanisms to determine message credibility, the exploitation of social media is garnering a significant amount of attention in contemporary research including the IS discipline's most highly regarded publications. A recent MIS Quarterly article (Bharadwaj, Sawy, Pavlou and Venkatraman, 2013) incorporated social media as a component of every theme in their proposed framework intended to guide digital business strategy development leading to the next generation of insights. This recognition of social media as a significant trend was due to the pervasiveness, rapidity of information flow and network effects provided by social media and forms as a component of every single aspect (scope, scale, speed and source) of their framework (Bharadwaj et al., 2013). Another example of a recent article, this time from Information Systems Research, provides an empirical assessment of consumer's use of social media in decision making. Goh, Heng and Lin (2013) combined Facebook with a customer reward program database to ascertain the positive effects of information dissemination and information seeking through social media upon consumer purchasing decisions.

It necessity for practitioners and academics alike to incorporate and exploit social media data and effects into their modeling and theoretical frameworks is readily apparent. It is with this motivation that we provide our argument for an assumption of credibility relating to at least some social media services. We next describe the mechanism with which we intend to collect, measure and assess these messages for use in assessing the dimensions of culture. Our intent is that this description serves not only our purpose, but to also describe an extensible framework by which researchers can become familiar with the collection and assessment techniques available to them, and then expand upon it to provide unique solution for their own inquiries.

METHODOLOGY

It is important to recognize that data collection from social media is simply another mechanism by which researchers can acquire data relating to their focal questions and topics of inquiry. Using social media and big data do not provide any mechanism by which researchers can eschew theoretical development. Indeed, our intent is simply to provide an alternative source for data usable by the traditional statistical methods relating to path analysis. Therefore, as with all academic inquiries, we must first describe the model. This includes both the measurement and structural components. Given the confirmatory and demonstrative nature of this article, we selected a set of theoretical constructs with a rich history of rigorous inquiry and iterative development and confirmation across time. We use several of the cultural dimensions described in the current Values Survey Module 2013 (VSM 2013) (Hofstede et al., 2010).

We selected this set of factors due to their nature of spanning multiple conceptual areas with little to no overlap. We also selected them due to their ability to exemplify the process of defining, capturing, measuring and analyzing multiple distinct factors. Although the intention is that ultimately the results of this method are usable in statistical path analysis, the dimensions of culture are not causally linked and so require no path analysis. This furthers the goals of this research by simplifying the outcomes as a set of factors useful for creating a comparative index without necessitating statistically significant causal linkages at this point in time. But at the same time, we intend to extend the extant boundaries of exploitation of social media and research of cultural dimensions beyond its current limited state; for example comparing emoticon usage in Twitter messages across cultures based on the cultural dimension indices (Park, Baek and Cha, 2014).

The dimensions of culture are currently comprised of a set of six factors (Power Distance, Individualism v/s Collectivism, Masculinity v/s Femininity, Uncertainty Avoidance, Long v/s Short Term Orientation, Indulgence v/s Restraint) with which we need to provide a method to measure them. In order to develop a measurement instrument suitable for social media data, we refer to the original measurement instrument. The current version for measuring the cultural index is the VSM 2013 and is comprised of 30 questions of which five relate to demographic data. These questions will form the basis for the development of our instrument, but require significant alteration from their original state.

Due to space considerations for this article, we are limiting the inquiry to a single cross-sectional data collection and assessment mechanism that we can use to derive cultural index values for comparative purposes. We will further limit the inquiry to two cultural dimensions well suited to a cross sectional inquiry. Specifically, we select Long v/s Short Term Orientation (L/STO) and Indulgence v/s Restraint (I/R) for assessment by this inquiry. These two dimensions are associated with questions 11, 12, 13, 14, 16, 17, 19 and 22 from the original survey (see Table 2 below). The suitableness of these questions is based upon the ability to isolate specific keywords and concepts that can be used to define a set of relevant data items to capture at a single point in time for later assessment without necessitating the measurement of deltas or changes over time. For example, we can isolate the concept of "important", "free time" and "fun" from the question text associated with item 11. Refer to Table 2 for the examples of extracting keywords from the original item verbiage.

Question # (Factor)	**Item text**	**Keywords**
11 (I/R)	How important is keeping time free for fun?	"important", "free time", "fun"
12 (I/R)	How important is moderation: having few desires?	"important", "moderation", "few desires"
13 (L/STO)	How important is doing a service to a friend?	"important", "doing service", "friend"
14 (L/STO)	How important is thrift?	"important", "thrift"
16 (I/R)	Are you a happy person?	"happy", "you"
17 (I/R)	Do other people or circumstances ever prevent you from doing what you really want to?	"people or circumstances", "prevent", "doing what you want"
19 (L/STO)	How proud are you to be a citizen of your country?	"pride", "citizenship"
22 (L/STO)	Persistent efforts are the surest way to results.	"persistence", "leads to", "results"

Table 2. Selected VSM 2013 Items (Hofstede et al., 2010)

The Twitter social media service provides a set of open Application Programming Interfaces (API), known as the REST and Streaming APIs (https://dev.twitter.com/rest/public), by which anyone may acquire messages posted to Twitter. A researcher can acquire a random subset of all tweets in real-time, or she can specify a set of keywords that will provide a message stream pertinent to the researcher's needs. The specified keywords limit the scope of the returned data to only those messages containing the keywords associated with the query. This is the reason we previously distilled the original item verbiage into keywords. The API queries can be accomplished via custom programs created in any modern programming language (Java, C#, Python, etc.) and Twitter returns the data in JSON format that can be stored natively by many database engines or as plain text for those that do not support native JSON data types.

But this only provides the raw data based upon our keywords that are hopefully relevant to our intended concepts. Next, we must perform a careful cleaning of the data. Automated classification can assist with this task; automated classifiers have been successfully used for many years to detect Unsolicited Commercial Email (UCE or "spam") messages. This is similar to what needs to be cleaned from the raw data, messages completely unrelated to our focal concept but that happen to contain our specified keywords.

Naieve Bayesian classifiers are provided in many modern text processing packages such as Natural Language Toolkit (NLTK) (http://www.nltk.org). These premade classification routines eliminate the tedious chore of creating, testing and validating custom coded classification routines. However, although these classifiers are generic in usage; they still must be trained. This is accomplished by a process similar to using hold-back data to train and validate statistical models. In this instance, it necessitates providing the classifier with a representative set of messages with their classification coded by the researcher and then validating the classifier with previously unseen data to test accuracy and reliability of the classifications.

This process allows us to inform the classifier that example tweet #1 (See Table 3 below) is important while example tweet #2 is not. The first is important because it includes a value statement, "love", regarding thrift which aligns precisely with question 14 from the VSM 2013. Although shopping at thrift stores is only a single facet of the concept of thrift, this is one activity by which thrift will manifest in a thrifty person. The second example in contrast, only tells us that the author buys books but doesn't include any information relating to a valence or magnitude of the author's value towards thrift. Therefore, it would be appropriate to include this first statement in our analysis of cultural values towards thrift and exclude the latter.

Example Tweet #1	Example Tweet #2	Example Tweet #3
"i love shopping at thrift shops"	"I can't go thrift shopping without buying books, it just doesn't happen"	"At first I hated Thrift shop but now I'm gonna pop some tags, Only got twenty dollars in my pocket."

Table 3. Example Tweets Captured December 13, 2014

The results of our classification routines provide us with the raw data that still needs a quantitative measure in order to compute the cultural indices. The original instrument provided a five-point scale for the respondent to indicate the degree of importance or frequency of occurrence for the appropriate item. Therefore, we must identify some mechanism by which the message contents can be expressed as a quantitative measure similar to the original 5-point scale. Sentiment analysis can provide such a measure of message content and has been used in prior academic research. For example, Ringsquandl and Petkovic (2013) used sentiment analysis to ascertain trends in statement sentiment for Twitter messages regarding Republican Party candidates for U.S. President during specific time periods of the 2012 primaries.

Sentiment scores provide the ability to ascertain the valence as well as the magnitude of the overall sentiment of the message. Another open source package, Pattern (DeSmedt and Daelemans, 2012), provides routines for sentiment assessment that returns values on a scale from -1 to +1 based upon the message contents (word usage, emoticons, punctuation, etc.). Given the 140 character limit Twitter places on their messages, tweets are generally limited to a single topic as there simply is not enough space for complex thoughts or phrases spanning multiple content domains. Therefore, we contend that the sentiment measure of a tweet is an appropriate proxy for the value the author places upon the tweet topic for the majority of messages on Twitter. Messages where such an assumption may not be appropriate can be eliminated at the data cleaning stage, filtered by the automated classifier, or retained in the dataset as outliers.

Lastly, we must account for the complex messages as shown in Tweet #3 (see Table 3 above). A human reader can readily interpret this as a positive statement relating to thrift. However, there are several problems that must be overcome in order to process this message via automatic classification and sentiment analysis. First, is the sentiment change over time provided by the "at first" and "but now" message contents. Second, is the usage of negative keywords "hated" that will negatively influence the sentiment scoring even though the author's intent is to profess a positive sentiment relating to thrift shopping. This is similar in effect to sarcasm which is difficult for an automated agent to appropriately account for. Lastly, is the problem of typographical errors (intentional or accidental), lexical variations in word usage, acronyms and abbreviations. Given the vast amounts of data collectible via social media streams, messages such as these are likely best handled by sorting into categories for later analysis by humans. This allows researchers to focus their attention on the problematic messages until such a time as enough examples are gathered so as to train a classifier in the appropriate ways to handle such instances.

LIMITATIONS AND CONCLUSION

A key limitation of this research is that our contention of credibility expectation is based upon theory and not derived from empirical evidence. Although our contentions share similarity with the distributional expectations used in probability theory relating to the central limit theorem and law of large numbers, an empirical inquiry of the distributional characteristics of credibility in social media messages could greatly enhance our contentions and serve as a useful extension of this research. A second limitation of this research is that we only describe and use a cross-sectional data collection routine. Although a more complex collection design and routine would be necessary to monitor and measure various factors over time, the actual collection mechanisms will be similar to that described here. Therefore, we expect researchers to be able to use this cross-sectional example as a generic template for more complex data collection such as time-series or panel data.

Social media research not only enhances the relevancy of IS in the post web 2.0 age, but also provides an opportunity for undergraduate students and research assistants to directly contribute to research projects by using the technical skills acquired during their IS programs. This means that there is no necessity, and this article does not intend to propose, that IS researchers become immersed in the technical intricacies of web services, social media protocols, cloud technologies, etc. This can instead serve as the contribution of the research assistant. Additionally, there is also the potential for tremendous benefits as it opens an entirely new avenue of data collection independent of the difficulties (carelessness, common method bias, etc.) associated with traditional data collection methods such as surveys. The opportunities provided by social media data streams create a tremendous opportunity for IS research to answer the often identified disconnect between academics and practitioners in IS research (Benbasat and Zmud, 1999). By exploiting user generated content such as social media streams in conjunction with decades of rigorous theory development, IS research stands at the cusp of providing significant new contributions to the ever increasing necessity of modern businesses to acquire and exploit new sources of actionable insights.

REFERENCES

1. Benbasat, I., and Zmud, R. W., (1999). Empirical research in information systems: The practice of relevance. MIS Quarterly, Vol. 23(1), pp. 3-16.
2. Bharadwaj, A., Sawy, O., Pavlou, P., and Venkatraman, N. (2013). Digital business strategy: Toward a next generation of insights. *MIS Quarterly*, 37, 2, 471-482.
3. Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.
4. Crockford, D. (2006). JSON: The fat-free alternative to XML, In *Proceedings of XML*. Boston, Massachusetts, Vol. 2006.
5. Curtis, L., Edwards, C., Fraser, K., Gudelsky, S., Holmquist, J., Thornton, K., and Sweetser, K. (2009). Adoption of social media for public relations by nonprofit organizations. *Public Relations Review*, 36, 90-92.
6. DeSmedt, T., and Daelemans, W. (2012). Pattern for python, *Journal of Machine Learning Research*, 13, 2031-2035.
7. Federal Rules of Evidence. Rule 803. Exceptions to the rule against hearsay. Retrieved 06 October, 2014 from http://www.law.cornell.edu/rules/fre/rule_803
8. Fielding, R. T. (2000). *Architectural styles and the design of network-based software architectures*. (Doctoral dissertation, University of California, Irvine)
9. Goh, K., Heng, C., and Lin, Z. (2013). Social media brand community and consumer behavior: Quantifying the relative impact of user – and marketer – generated content. *Information Systems Research*, 24, 1, 88-107.
10. Hofstede, G., Hofstede, G., and Minkov, M. (2010) *Cultures and Organizations: Software of the Mind. Revised and Expanded Third Edition*. New York: McGraw Hill.
11. Krikorian, R. (2013). New Tweets per second record, and how!, Twitter Inc. *Engineering Blog*, Retrieved 05 October, 2014 from https://blog.twitter.com/2013/new-tweets-per-second-record-and-how
12. Laney, D. (2001). 3D data management: Controlling data volume, velocity, and variety. *Application Delivery Strategies, File 949*, META Group.
13. Longley-Cook, L. H. (1962). *An Introduction to Credibility Theory*, PCAS, Vol. 49, pp. 194-221.
14. Park, J., Baek, Y., and Cha, M. (2014). Cross-cultural comparison of nonverbal cues in emoticons on Twitter: Evidence from big data analysis. *Journal of Communication*, 64, 333-354.
15. Paul, M. J., and Dredze, M. (2011). You are what you tweet: Analyzing Twitter for public health. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Niagara Falls, Ontario, Canada.
16. Sussman, S. W., and Siegal, W. S. (2003). Informational influence in organizations: An integrated approach to knowledge adoption. *Information Systems Research*, 14, 1, 47-65.
17. Ringsquandl, M., and Petkovic, D., (2013). Analyzing political sentiment on Twitter. *Proceedings of the 2013 AAAI Spring Symposium*, Palo Alto, CA, USA.
18. Rynes, A., and Bjornard, T. (2011). *Intent, capability and opportunity: A holistic approach to addressing proliferation as a risk management issue*. Idaho National Laboratory, U.S. Department of Energy.
19. Twitter, Inc. (2014). About Twitter, Inc., Retrieved on 07 December, 2014 from https://about.twitter.com/company

20. Venkatesh, V., Morris, M., Davis, G., and Davis, F., (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27, 3, 425-478.
21. YouTube Inc. (2014). Press Room - Statistics, Retrieved on 07 December, 2014 from https://www.youtube.com/yt /press/statistics.html